Predictive Modeling: Principles and Practices

Rick Hefner, Dean Caccavo
Northrop Grumman Corporation

Philip Paul, Rasheed Baqai
Unlimited Innovation, Inc.

NDIA Systems Engineering Conference
20-23 October 2008

Background

- **Predictive modeling relies on historical program performance data (predictive analytics) in conjunction with a forecasting algorithm model to predict future outcomes**
  - Ranges from simple extrapolation techniques to sophisticated Neural Network based models
- **This presentation will discuss the principles of predictive modeling, outline the fundamental methods and tools, and present typical results from applying these techniques to project performance**

Agenda

- ✓ **What is Predictive Analysis?**
- Recent Trends
- Application to Program Performance
- Pilot Results and Feedback
- Summary

What is Predictive Analysis?

- ***Could this network packet be from a virus attack?***
  - Predict likelihood of the network packet pattern
  - ➔ **Anomaly detection (outlier detection)**
  - Similar questions:
    - Are the hospital lab results normal (Adverse drug effect detection)
    - Is this credit transaction fraudulent? (fraud detection)
- ***Will this student go to college?***
  - Based on Gender, ParentIncome, ParentEncouragement, IQ, etc.
  - E.g., if ParentEncouragement=Yes and IQ>100, College=Yes
  - ➔ **Classification (prediction)**
  - Similar questions:
    - Is this a spam email? (spam filtering)
    - Recognition of hand-written letters (pen recognition)
- ***What is the person’s age?***
  - Based on Hobby, MaritalStatus, NumberOfChildren, Income, HouseOwnership, NumberOfCars, …
  - E.g., If MaritalStatus=Yes, Age = 20+4*NumberOfChildren+0.0001*Income+…
  - ➔ **Regression (prediction)**

Agenda

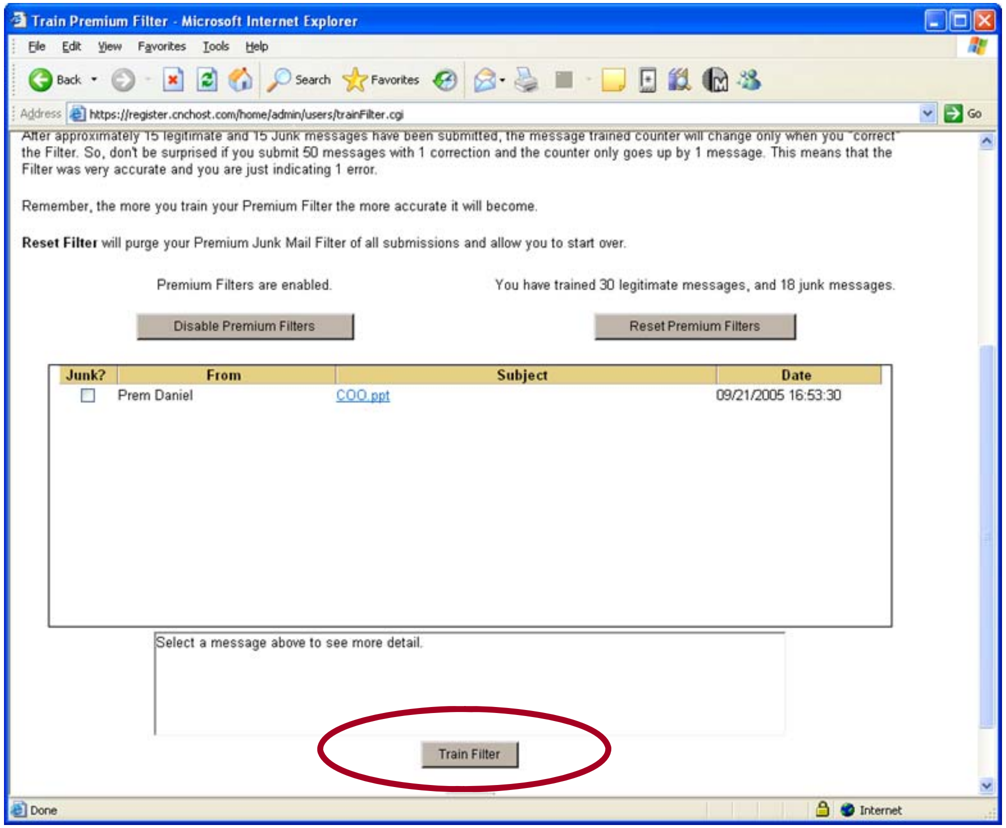
- ▪ What is Predictive Analysis?
- ✓ **Recent Trends**
- ▪ Application to Program Performance
- ▪ Pilot Results and Feedback
- ▪ Summary

Predictive Analysis Trends – Adoption is on the rise

- Predictive Analysis is becoming more prevalent and integrated in business applications
  - o *Example: Disease management and evidence based care, based on historical diagnosis and procedure codes of patients*
  - o *Example: E-Mail filtering using predictive analysis*
- Predictive Analysis algorithms are being integrated into existing databases, data mining tools
  - o *Example: Microsoft SQL Server 2005 has predictive analysis algorithms*

Example:

Premium predictive analysis based filtering on e-mail, available to any e-mail user



After approximately 15 legitimate and 15 Junk messages have been submitted, the message trained counter will change only when you "correct" the Filter. So, don't be surprised if you submit 50 messages with 1 correction and the counter only goes up by 1 message. This means that the Filter was very accurate and you are just indicating 1 error.

Remember, the more you train your Premium Filter the more accurate it will become.

Reset Filter will purge your Premium Junk Mail Filter of all submissions and allow you to start over.

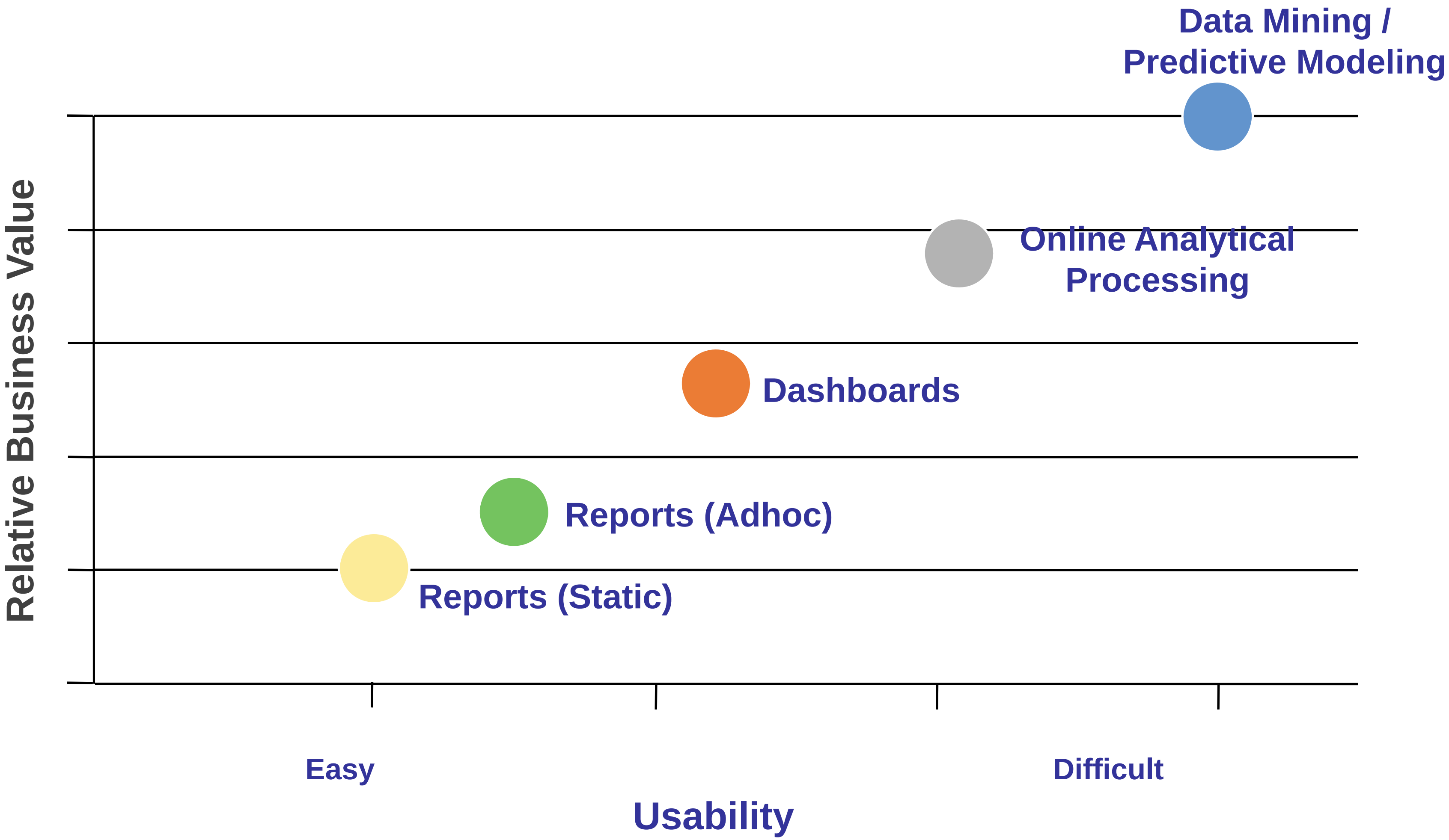
Premium Filters are enabled.

You have trained 30 legitimate messages, and 18 junk messages.

Junk?	From	Subject	Date
□	Prem Daniel	COO.ppt	09/21/2005 16:53:30

Select a message above to see more detail.

Predictive Analysis Trends – Tools are becoming easier to use



Predictive Analysis Trends – Model development is more structured

Define a Model

Train the Model

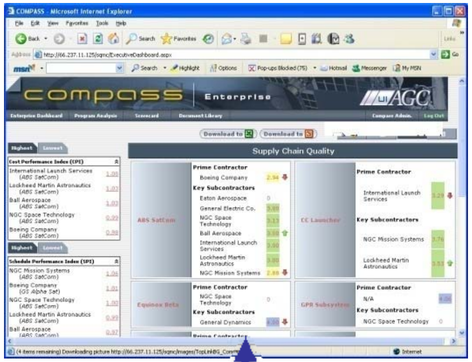
Training Data

Test the Model

Test Data

Prediction using the Model

Prediction Input Data



COMPASS - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://66.237.11.125/sqmc/ExecutiveDashboard.aspx Go
msn Search Highlight Options Pop-ups Blocked (75) Hotmail Messenger My MSN
compass Enterprise uiAGC
Enterprise Dashboard Program Analysis Scorecard Document Library Compass Admin. Log Out
Download to Download to
Highest Lowest

Cost Performance Index (CPI)

International Launch Services (ABS SatCom)	1.08
Lockheed Martin Astronautics (ABS SatCom)	1.03
Ball Aerospace (ABS SatCom)	1.03
NGC Space Technology (ABS SatCom)	0.99
Boeing Company (ABS SatCom)	0.98

Schedule Performance Index (SPI)

NGC Mission Systems (ABS SatCom)	1.06
Boeing Company (GS Alpha Sat)	1.01
NGC Space Technology (ABS SatCom)	1.00
Lockheed Martin Astronautics (ABS SatCom)	0.99
Ball Aerospace (ABS SatCom)	0.97

Supply Chain Quality

ABS SatCom
Prime Contractor
Boeing Company 2.94
Key Subcontractors
Eaton Aerospace 0
General Electric Co. 3.88
NGC Space Technology 3.13
Ball Aerospace 3.00
International Launch Services 3.00
Lockheed Martin Astronautics 3.00
NGC Mission Systems 2.88

CC Launcher
Prime Contractor
International Launch Services 3.29
Key Subcontractors
NGC Mission Systems 3.76
Lockheed Martin Astronautics 3.53

Equinox Beta
Prime Contractor
NGC Space Technology 0
Key Subcontractors
General Dynamics 4.00

GPR Subsystem
Prime Contractor
N/A 4.06
Key Subcontractors
NGC Space Technology 0

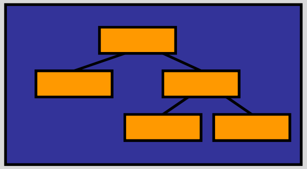
Prime Contractor

(4 items remaining) Downloading picture http://66.237.11.125/sqmc/images/TopLinkBG_Com... Internet

Off-the-shelf or Proprietary Predictive Analysis Engine

Third Party Predictive Analysis tools

Off-the-shelf or proprietary Predictive Analysis Model



- Executive understanding of the creation, training and testing of the model is critical to success
- The Model gets more powerful and accurate as the volume of data fed into the model increases

Predictive Analysis Trends – Algorithms are available for use

Decision Trees	Naïve Bayesian	Clustering	Sequential Clustering	Time Series	Association rules	Neural Network	
1	2	2	2		2	1	Classification
1	2	2	2			1	Regression
		1	1			2	Segmentation
1	1	2	2		1	1	Association Analysis
		1	1			2	Anomaly Detect.
			1				Sequential Analysis
				1			Time series

1 - First Choice **2 - Second Choice**

Data Mining Vendors & Tools

- SAS (Enterprise Miner)
- IBM (DB2 Intelligent Miner)
- Oracle (ODM option to Oracle 10g)
- SPSS (Clementine)
- Insightful (Insightful Miner)
- KXEN (Analytic Framework)
- Prudsys (Discoverer and its family)
- Microsoft (SQL Server 2005)
- Angoss (KnowledgeServer and its family)
- DBMiner (DBMiner)
- Many others

Agenda

- What is Predictive Analysis?
- Recent Trends
- ✓ **Application to Program Performance**
- Pilot Results and Feedback
- Summary

Mission Assurance Continuum

	Program Performance Oversight	Program Analysis Reporting	Predictive Program Health
	Industry Minimum	**Industry Best Practice**	**Industry Innovators**
	Proactive Program Management Program Portfolio Management	**Reports based on current and passed performance data of portfolio programs, programs, and subcontract reports**	**Predictive Analysis based on Program Performance Modeling**
Approach and Scope	• Self reported Program Portfolio includes critical and high visibility programs • Standard Program Management Metrics collected on a periodic basis	• Self Reported Program metrics collected periodically and at specific program milestones • Reporting analysis performed as needed	• Self reported program metrics, organizational data, personnel data and customer reported metrics collected at regular intervals • Predictive models developed using historical data (leading indicators rationalized) • Models validated against historical data
Infrastructure and Breadth	• Program data maintained by individual programs • Summary information provided to enterprise repository	• Program data collected periodically into an enterprise-wide program management repository • Program, Enterprise and Subcontracts performance reports available	• Holistic enterprise wide approach to program execution • Models continually refined using current program performance data • Sophisticated predictive measures provided to programs and enterprise
Data Requirements	• Very few metrics collected from programs • Key program metrics (cost performance, schedule performance, technical performance, CPI, SPI etc.) • Standardized program taxonomy information like customer, contract type	• 25 – 100 metrics collected from programs • Key program metrics collected at all specified Program Milestones.	• 50 – 75 metrics collected from programs and refined to include only the few relevant metrics • Adaptive approach to qualitative and quantitative performance indicators • Direct and Indirect metrics collected for the programs; qualitative information is mined • Proactive responses based on predictive analysis of ongoing and historical performance

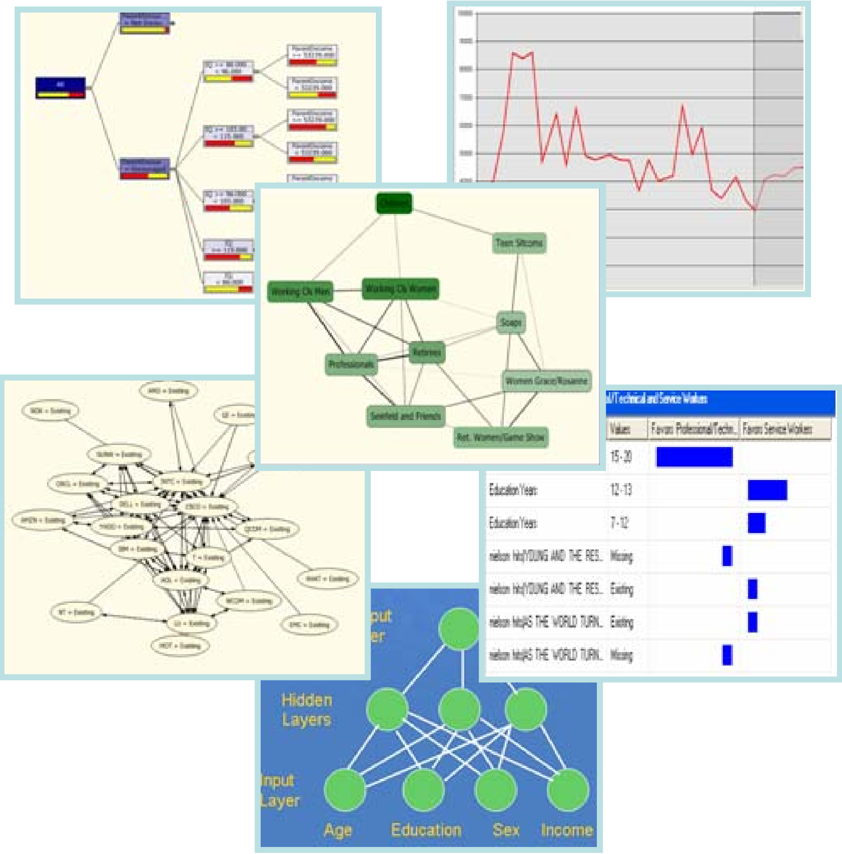
Overarching Objectives for Predictive Modeling

- Provide program management staff with Predictive Models to "test-their-gut" against enterprise experience data before making strategic program decisions
- Develop Predictive Models that provide insight into identifying "headlight metrics" that influence Schedule and Cost realism during program execution
- Leverage existing enterprise information to develop Predictive Models for programs
- Ensure that models are extensible and automatically calibrated with additional data from the program and enterprise

Potential Areas for Predictive Analysis

Potential Predictive Analysis Models for Program Management and Subcontractor Management

- Schedule Risk at WBS level based on past performance
- Cost Risk at WBS level based on past performance
- Technical Risk at WBS level based on past performance
- Spending and staffing profile for the program life cycle
- Subcontractor risk profile based on past performance
- Sub-tier quality at subcontract and WBS level
- Defect/Aberrations for the program life cycle
- Mission Assurance models based on program category



Predictive Analysis Algorithms

- Decision Trees
- Naïve Bayesian
- Clustering
- Sequence Clustering
- Association Rules
- Neural Network
- Time Series
- Custom Model

Predictive Analysis High Level CONOPS

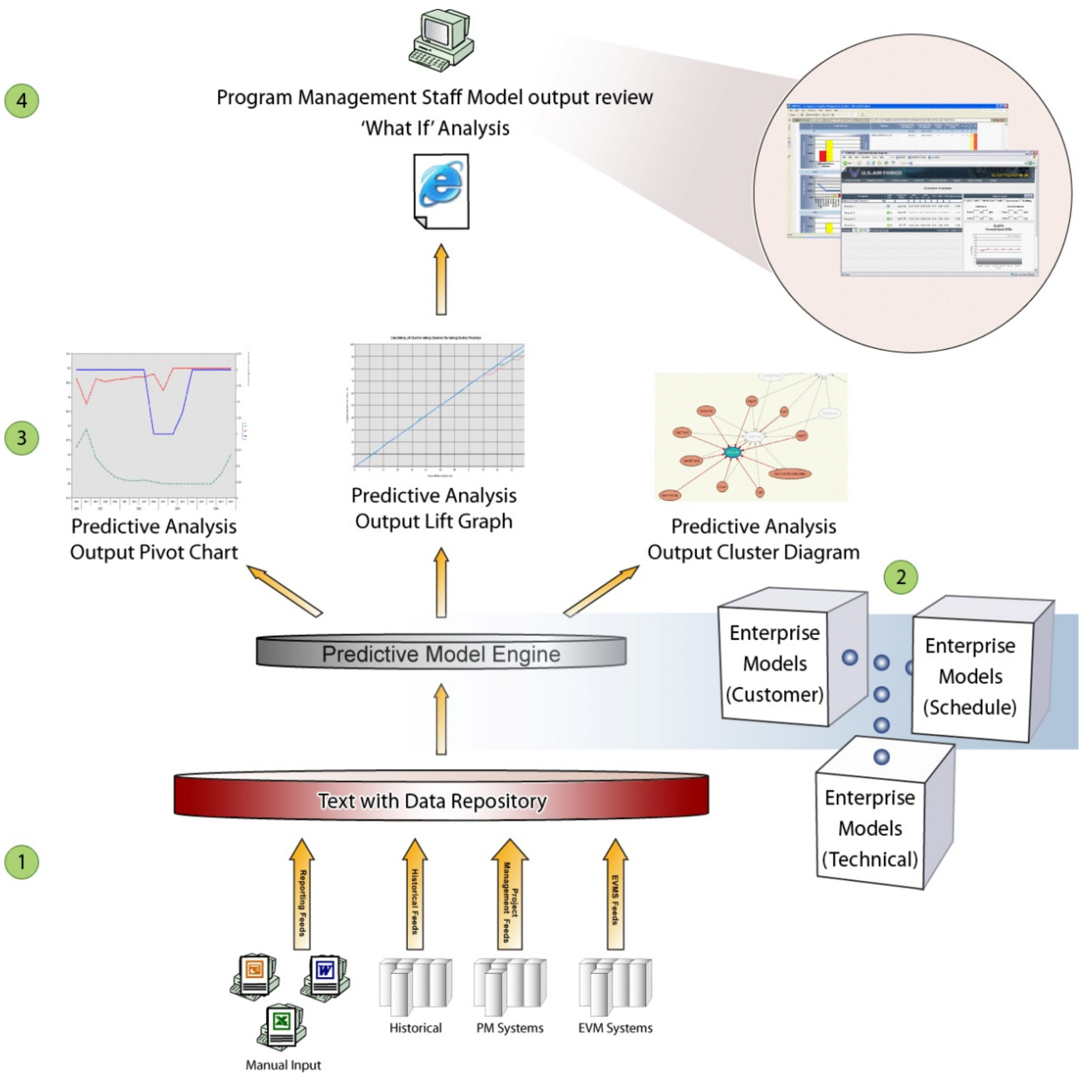
1) Enterprise data is mined and analyzed

2) Enterprise models are defined by Analysts

3) Enterprise model outputs are defined by Analysts and customized by PM staff

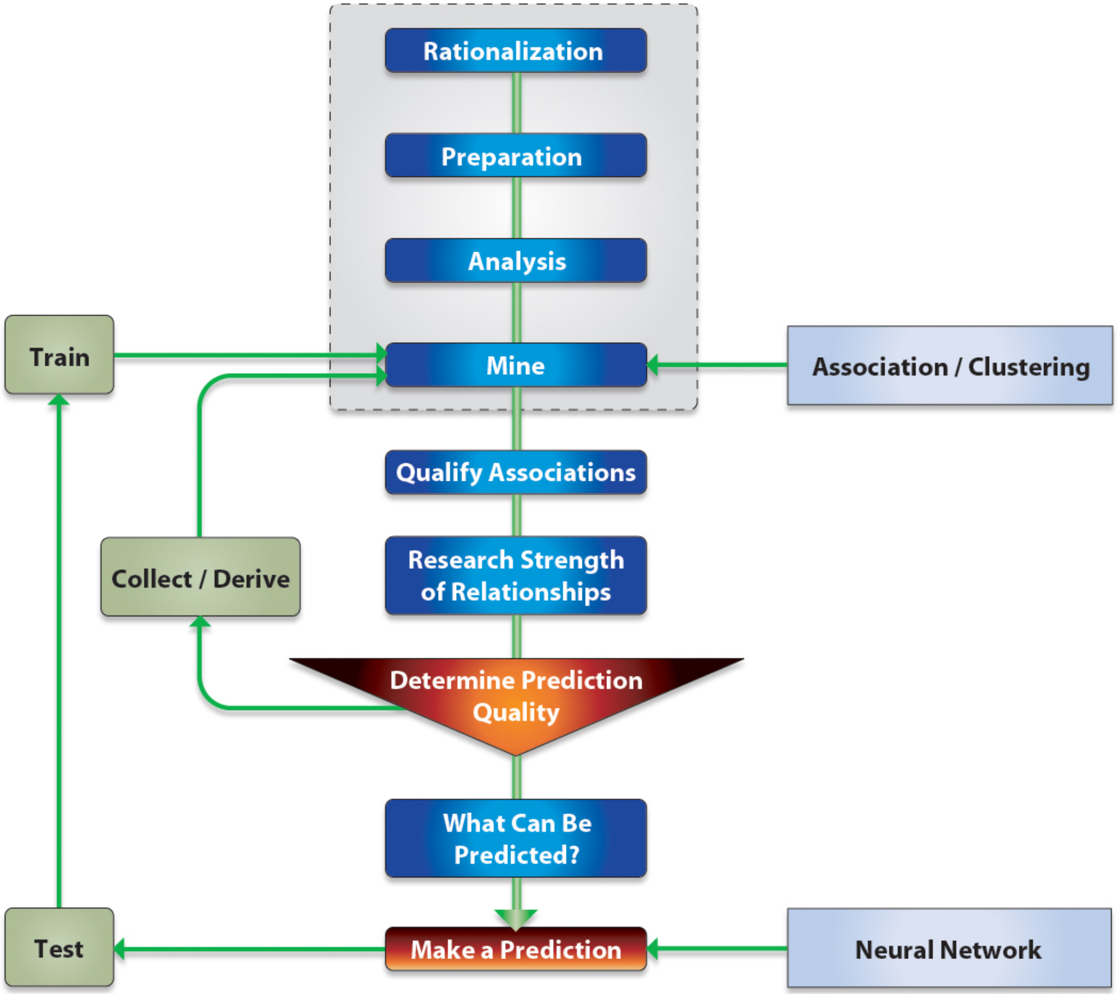
4) PM staff use models interactively

Key Benefit:
Leverages enterprise experience data and sophisticated algorithms into predictive models for cost and schedule realism checks during program execution



The Predictive Modeling Process

- Explore the Data
- Understand Data Relationships
- Derive/Enhance the Data
- Use the Data to Predict
- Train the Model



What can be Predicted with Reasonable Accuracy?

Program Lifecycle Stage	Limited Number of Programs	Enterprise Experience
Large volume of historical data	▪ Likelihood or return to acceptable performance ▪ Predictive Program Performance 1	▪ Quadrant 2 predictions ▪ Quadrant 3 predictions ▪ Early warning “headlight indicators” ▪ Higher accuracy based on enterprise experience 3
Limited Historical data		▪ Cost, schedule realism ▪ Phase realism ▪ WBS Accuracy 2
	Low	High

Volume of “Like” Programs

Agenda

- Background
- Industry Trends
- Application to Program Performance
- ✓ **Pilot Results and Feedback**
- Summary

Predictive Modeling Pilot Objectives

- **Provide program management staff with Predictive Models to "test-their-gut" against enterprise experience data before making strategic program decisions**
- **Develop Predictive Models that provide insight into identifying "headlight metrics" that influence Schedule and Cost realism during program execution**
- **Leverage existing enterprise information to develop Predictive Models for programs**
- **Ensure that models are extensible and automatically calibrated with additional data from the program and enterprise**

Pilot Approach

- Analyze and rationalize the available enterprise data
  - Enterprise Level Office of Cost Estimation and Risk Assessment (OCERA) data
  - Division Level Stoplight Program data
  - Program Level Program Review Authority (PRA) data for relevant programs
- Develop predictive modeling approach to provide schedule and cost measures during program execution phase
- Develop preliminary predictive models using appropriate algorithms and mining existing enterprise data
  - Mining – Clustering, Decision Trees and Naïve Bayesian Algorithms
  - Predictions – Neural Network, Bayesian Algorithms and Clustering
- Get Pilot participation from three representative program types:
  - Large Scale System Integration Low Rate Initial Production program
  - Medium Sized Software program
  - Small IT System (Software and Hardware) program

Key Benefit: Leverages enterprise experience data and sophisticated algorithms into predictive models for use during program execution

Data analyzed for developing preliminary models

Data	Stoplight	OCERA	PRA
Data Period	2.5 years	5 – 6 years	Past 4 months
Frequency	Quarterly/Some older data is monthly	Major milestones or annually	Monthly
Breadth and depth of data	Monthly snapshot of key metrics	Very deep, very broad, with significant contextual information	Very deep, mostly snapshot without significant contextual information
Approximate number of data elements	~ 20	~ 70 key attributes	~40 key attributes

Analyzed enterprise level (OCERA), division level (Stoplight) and program level (PRA) data

Some Actual Data Types Used to Develop Predictive Model Relationships

Program Data

- Contract Type
  - CPAF, FFP, CPFF
- Type of Program
- Period of Performance
- Number of Milestones
- Number of sub-contractors
  - Subcontract value
  - Subcontract performance
- Total Value
- Annual Sales
- Number of incremental deliveries
- Average staff count
- SPI, CPI
- EAC, BAC
- Number of EAC changes
- Number of ECR/ECP
- Defects
  - Injection by phase
  - Occurrence by phase
- Skills Data
- Program Review Data
- Project Initiation Review Data

Program Self Assessment

- Monthly Ratings
  - Schedule
  - Technical
  - Cost
  - Mission Assurance
  - Management
  - Process

External Data

- CPARS
- Customer satisfaction data
- Award Fees

Milestone Data

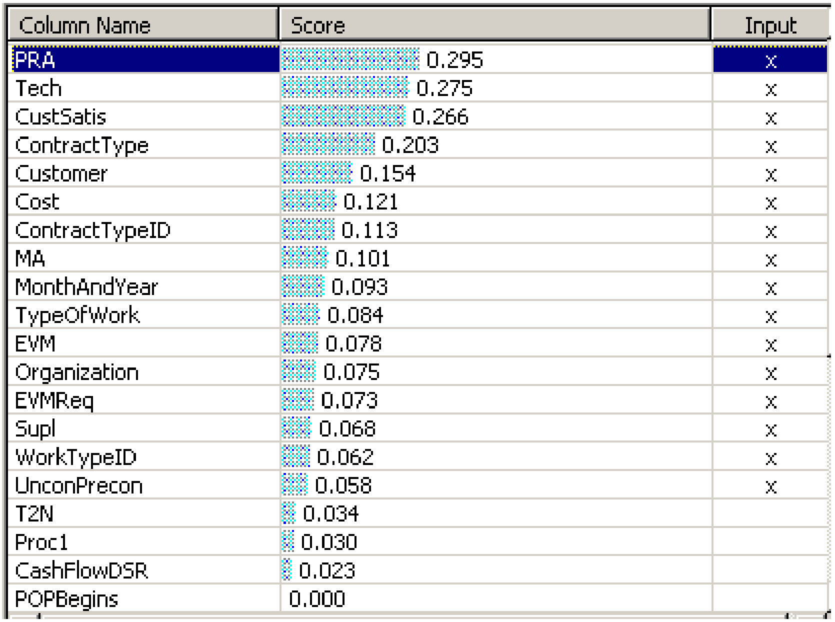
- Milestones
  - Proposal
  - Contract Startup
  - SRR
  - SDR
  - Software Specification Review
  - PDR
  - CDR
  - Test Readiness Review
  - Completion

Other Data

- Action Item Data
- Organization benchmark data
- SLOC, ESLOC
- Productivity
- Language, Component type, complexity,
- Reuse ratios
- Platform, environment

Contains Enterprise, Division and Program Data

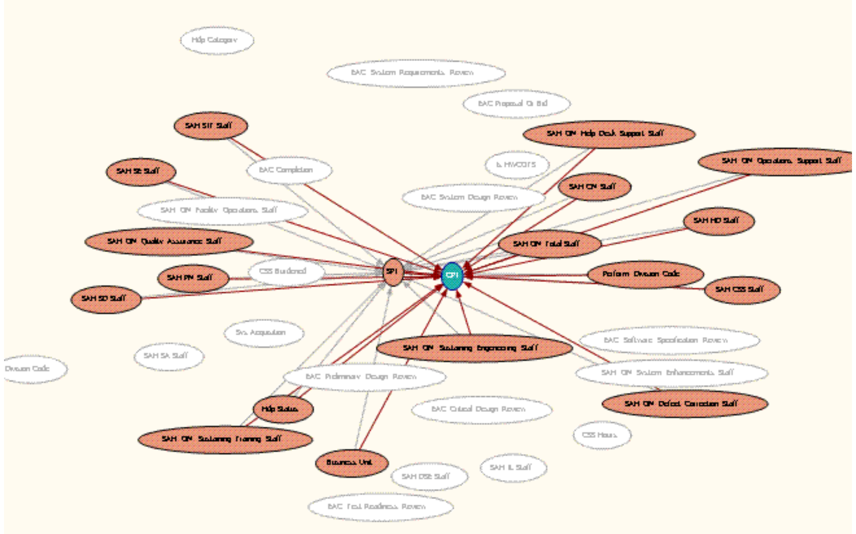
Data Mining Results



Column Name	Score	Input
PRA	0.295	x
Tech	0.275	x
CustSatis	0.266	x
ContractType	0.203	x
Customer	0.154	x
Cost	0.121	x
ContractTypeID	0.113	x
MA	0.101	x
MonthAndYear	0.093	x
TypeOfWork	0.084	x
EVM	0.078	x
Organization	0.075	x
EVMReq	0.073	x
Supl	0.068	x
WorkTypeID	0.062	x
UnconPrecon	0.058	x
T2N	0.034	
Proc1	0.030	
CashFlowDSR	0.023	
POPBegins	0.000	

Prediction Measures

- Schedule
- Cost


Hdp Category
BAC System Requirements Review
BAC Proposal Or Bid
SAH SIT Staff
SAH OM Help Desk Support Staff
SAH OM Operations Support Staff
SAH SE Staff
BAC Completion
& HWCOTS
SAH CM Staff
BAC System Design Review
SAH OM Facility Operations Staff
SAH HD Staff
SAH OM Quality Assurance Staff
SAH OM Total Staff
SPI
CPI
CSS Burdened
SAH PM Staff
Perform Division Code
SAH CSS Staff
SAH SD Staff
Sys Acquisition
BAC Software Specification Review
SAH SA Staff
SAH OM Sustaining Engineering Staff
Division Code
BAC Preliminary Design Review
SAH OM System Enhancements Staff
Hdp Status
BAC Critical Design Review
SAH OM Defect Correction Staff
SAH OM Sustaining Training Staff
CSS Hours
Business Unit
SAH DSE Staff
SAH IL Staff
BAC Test Readiness Review

- The mining showed that out of the over 125 metrics and measures some are leading indicators and are more important than others in influencing cost and schedule
- While it cannot be proved to be conclusive with the limited data that was used, the trends were definite

Derivation of Data & Data Relationships

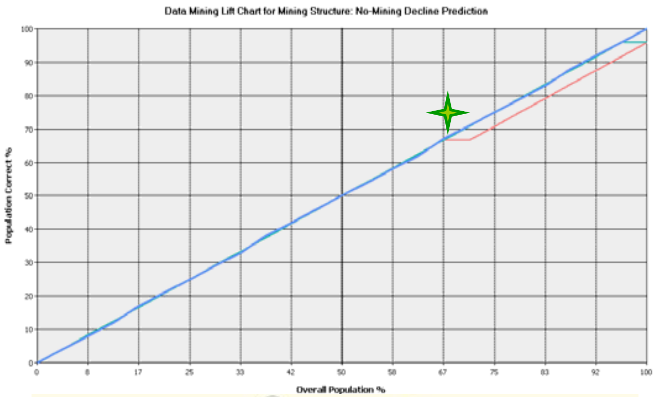
- **Examples of Derived Data**
  - Number of Outstanding Program Issues (with and without recovery dates)
  - Variance in program Cost/Schedule/Technical health from month-to-month
  - Program Cost/Schedule/Technical health trend from month-to-month
  - Variance in VAC from month-to-month taken as a percentage of the current EAC

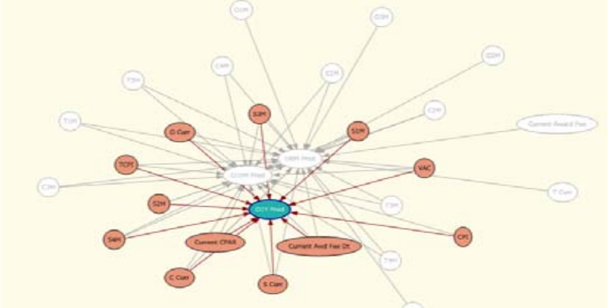
- **Examples of Discovered Relationships**
  - Schedule Health is a good indicator of program Overall Health recovery
  - Cost and Technical Health are good indicators of program Overall Health decline

Better understanding of the data allows for organization and enhancement of the dataset

Model Development & Calibration

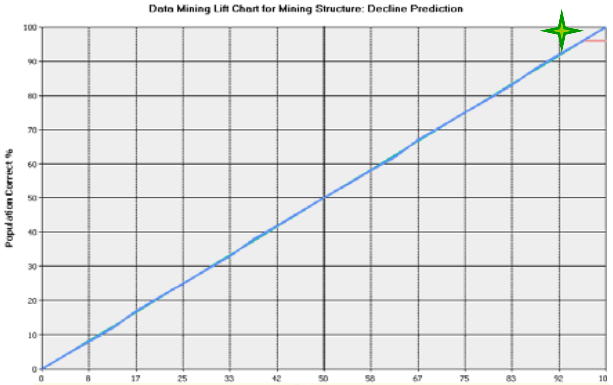
Model

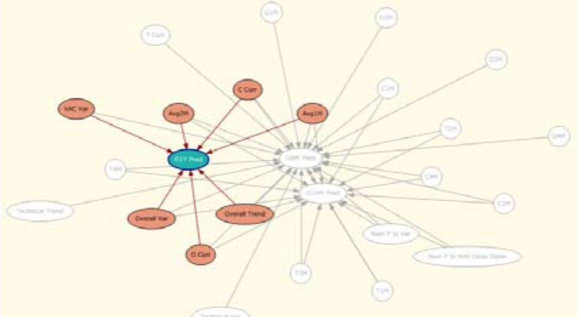




- Modeling without applied domain knowledge or calibration resulted in lower accuracy
- Association models able to determine relevant data attributes

Calibrated Model



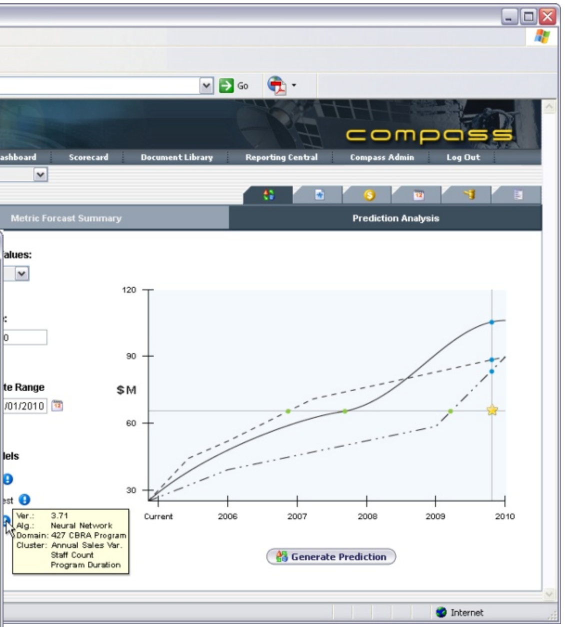


- Incorporating domain knowledge and calibration into data mining resulted in higher accuracy
- Data relationships are more clearly defined

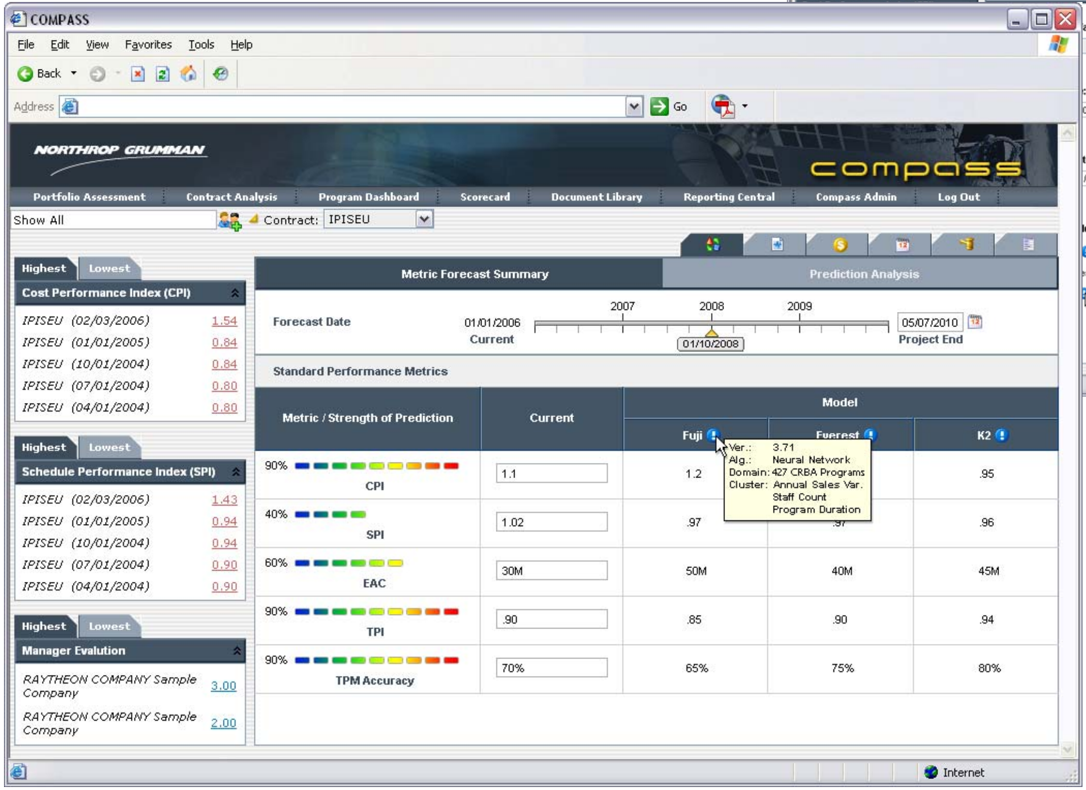
Domain knowledge & calibration applied to data mining can enhance the predictive model

Typical Results from the Models

Ability for Programs to review the predictive output from multiple models to “test-the-gut” before making strategic program decisions



Ver.: 3.71
Alg.: Neural Network
Domain: 427 CBRA Program
Cluster: Annual Sales Var.
Staff Count
Program Duration



Contract: IPISEU

Highest Lowest

Cost Performance Index (CPI)	
IPISEU (02/03/2006)	1.54
IPISEU (01/01/2005)	0.84
IPISEU (10/01/2004)	0.84
IPISEU (07/01/2004)	0.80
IPISEU (04/01/2004)	0.80

Highest Lowest

Schedule Performance Index (SPI)	
IPISEU (02/03/2006)	1.43
IPISEU (01/01/2005)	0.94
IPISEU (10/01/2004)	0.94
IPISEU (07/01/2004)	0.90
IPISEU (04/01/2004)	0.90

Highest Lowest

Manager Evalution	
RAYTHEON COMPANY Sample Company	3.00
RAYTHEON COMPANY Sample Company	2.00

Metric Forecast Summary | Prediction Analysis

Forecast Date: 01/01/2006 Current — 2007 — 2008 — 2009 — 01/10/2008 — 05/07/2010 Project End

Standard Performance Metrics

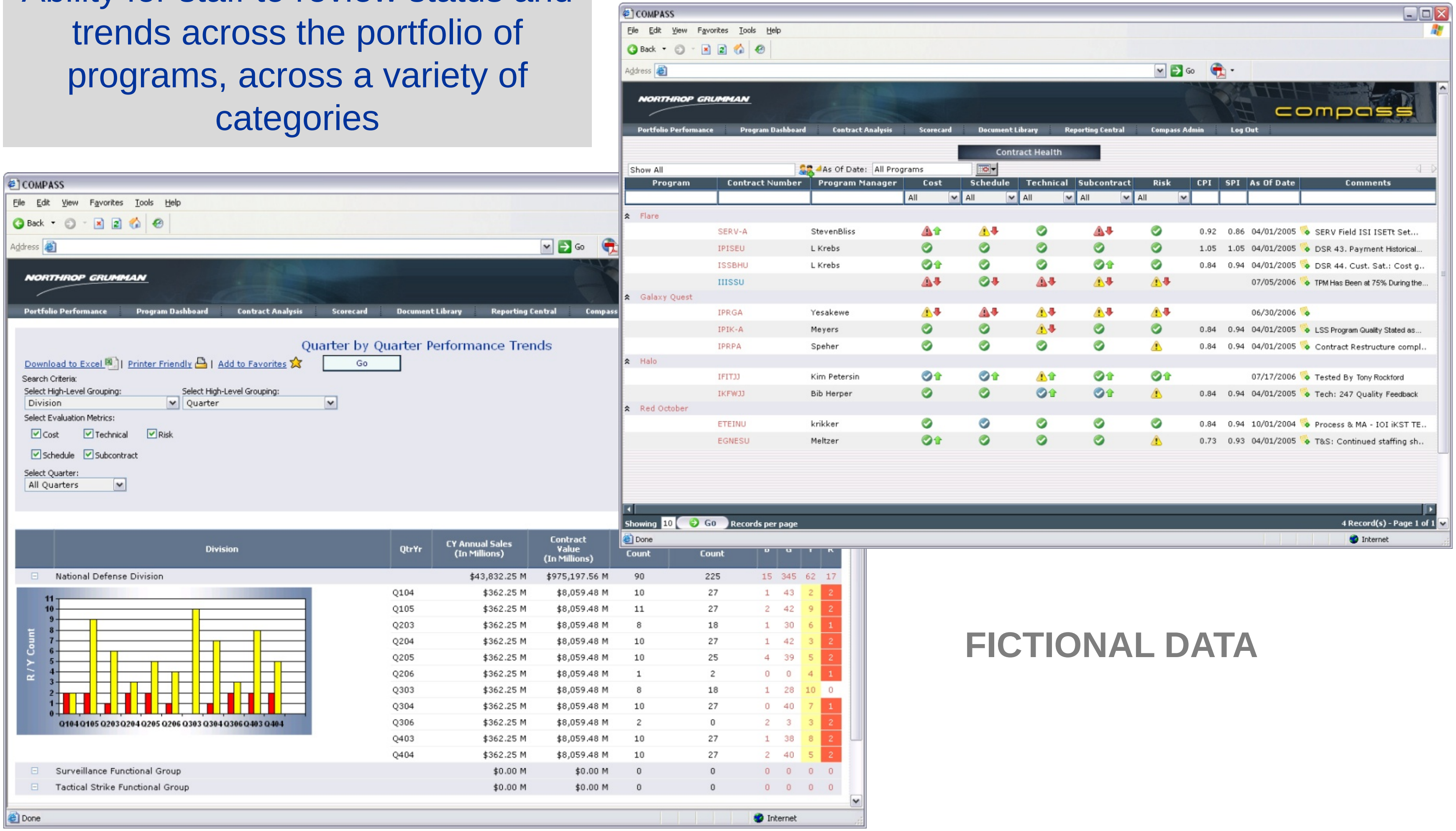
Metric / Strength of Prediction	Current	Model		
		Fuji	Everest	K2
90% CPI	1.1	1.2	[illegible]	.95
40% SPI	1.02	.97	.97	.96
60% EAC	30M	50M	40M	45M
90% TPI	.90	.85	.90	.94
90% TPM Accuracy	70%	65%	75%	80%

Ver.: 3.71
Alg.: Neural Network
Domain: 427 CRBA Programs
Cluster: Annual Sales Var.
Staff Count
Program Duration

FICTIONAL DATA

Typical Results from the Models

Ability for staff to review status and trends across the portfolio of programs, across a variety of categories



Contract Health

Show All · As Of Date: All Programs

Program	Contract Number	Program Manager	Cost	Schedule	Technical	Subcontract	Risk	CPI	SPI	As Of Date	Comments
			All	All	All	All	All				
Flare											
	SERV-A	StevenBliss	⚠↑	⚠↓	✓	⚠↓	✓	0.92	0.86	04/01/2005	SERV Field ISI ISETt Set...
	IPISEU	L Krebs	✓	✓	✓	✓	✓	1.05	1.05	04/01/2005	DSR 43. Payment Historical...
	ISSBHU	L Krebs	✓↑	✓	✓	✓↑	✓	0.84	0.94	04/01/2005	DSR 44. Cust. Sat.: Cost g..
	IIISSU		⚠↓	✓↓	⚠↓	⚠↓	⚠↓			07/05/2006	TPM Has Been at 75% During the...
Galaxy Quest											
	IPRGA	Yesakewe	⚠↓	⚠↓	⚠↓	⚠↓	⚠↓			06/30/2006	
	IPIK-A	Meyers	✓	✓	⚠↓	✓	✓	0.84	0.94	04/01/2005	LSS Program Quality Stated as...
	IPRPA	Speher	✓	✓	✓	✓	⚠	0.84	0.94	04/01/2005	Contract Restructure compl..
Halo											
	IFITJJ	Kim Petersin	✓↑	✓↑	⚠↑	✓↑	✓↑			07/17/2006	Tested By Tony Rockford
	IKFWJJ	Bib Herper	✓	✓	✓↑	✓↑	⚠	0.84	0.94	04/01/2005	Tech: 247 Quality Feedback
Red October											
	ETEINU	krikker	✓	✓	✓	✓	✓	0.84	0.94	10/01/2004	Process & MA - IOI iKST TE..
	EGNESU	Meltzer	✓↑	✓	✓	✓	⚠	0.73	0.93	04/01/2005	T&S: Continued staffing sh..

Showing 10 Go Records per page · 4 Record(s) - Page 1 of 1

Done · Internet

Quarter by Quarter Performance Trends

Download to Excel | Printer Friendly | Add to Favorites · Go

Search Criteria:

Select High-Level Grouping: Division

Select High-Level Grouping: Quarter

Select Evaluation Metrics:

☑ Cost ☑ Technical ☑ Risk

☑ Schedule ☑ Subcontract

Select Quarter: All Quarters

	Division	QtrYr	CY Annual Sales (In Millions)	Contract Value (In Millions)	Count	Count				
⊟	National Defense Division		$43,832.25 M	$975,197.56 M	90	225	15	345	62	17
		Q104	$362.25 M	$8,059.48 M	10	27	1	43	2	2
		Q105	$362.25 M	$8,059.48 M	11	27	2	42	9	2
		Q203	$362.25 M	$8,059.48 M	8	18	1	30	6	1
		Q204	$362.25 M	$8,059.48 M	10	27	1	42	3	2
		Q205	$362.25 M	$8,059.48 M	10	25	4	39	5	2
		Q206	$362.25 M	$8,059.48 M	1	2	0	0	4	1
		Q303	$362.25 M	$8,059.48 M	8	18	1	28	10	0
		Q304	$362.25 M	$8,059.48 M	10	27	0	40	7	1
		Q306	$362.25 M	$8,059.48 M	2	0	2	3	3	2
		Q403	$362.25 M	$8,059.48 M	10	27	1	38	8	2
		Q404	$362.25 M	$8,059.48 M	10	27	2	40	5	2
⊟	Surveillance Functional Group		$0.00 M	$0.00 M	0	0	0	0	0	0
⊟	Tactical Strike Functional Group		$0.00 M	$0.00 M	0	0	0	0	0	0

Done · Internet

FICTIONAL DATA

Agenda

- What is Predictive Analysis?
- Recent Trends
- Application to Program Performance
- ✓ **Summary**

Summary – Critical success factors

- Executive and Enterprise support and understanding of long-term strategic benefits
- Understanding of the types of data and the correlation between the data
- Understanding of the various constituents in the value chain and the tools/processes for each constituent
- Prototypes or mockups that depict the results of the model
- Sound and robust technical architecture
- Delivery mechanism that shields the complexity of the model from the end users

More Information

- OLE DB for DM specification
  - http://www.microsoft.com/downloads/details.aspx?FamilyID=01005f92-dba1-4fa4-8ba0-af6a19d30217&DisplayLang=en
- Plug-in
  - http://www.msnusers.com/AnalysisServicesDataMining/Documents/Files%2FSQL%20Server%20Data%20Mining%20Plug%2DIn%20Algorithms%20%28Beta%202%20%2B%2B%29.zip
  - A white paper, tutorial, and complete sample code for Pair-wise Linear Regression
- SQL Server 2005:
  - www.microsoft.com/sql/2005
- Community:
  - Microsoft.public.sqlserver.datamining
  - Microsoft.private.sqlserver2005.analysisservices.datamining
  - Groups.msn.com/AnalysisServicesDataMining
- msdn.microsoft.com (search "data mining")

- Decision trees (classification/regression):
  - ftp://ftp.research.microsoft.com/users/surajitc/icde99.pdf
  - http://www.research.microsoft.com/research/pubs/view.aspx?tr_id=81
  - http://research.microsoft.com/~dmax/publications/dmart-final.pdf
- Association rules:
  - Apriori algorithm (see Data Mining concepts and techniques)
- Clustering
  - EM:http://www.research.microsoft.com/scripts/pubs/view.asp?TR_ID=MSR-TR-98-35
  - K-means (see Data Mining concepts and techniques)
- Sequence clustering
  - ftp://ftp.research.microsoft.com/pub/tr/tr-2000-18.pdf
- Time series:
  - http://research.microsoft.com/~dmax/publications/dmart-final.pdf
- Neural network
  - Conjugate gradient method (see Data Mining concepts and techniques)
- Naïve Bayesian
  - See Data Mining concepts and techniques

Contact Information

Rick Hefner, Ph.D.
Northrop Grumman Corporation
(310) 812-7290
rick.hefner@ngc.com